

# Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*

Maha R Farhat<sup>1,28</sup>, B Jesse Shapiro<sup>2–5,28</sup>, Karen J Kieser<sup>6</sup>, Razvan Sultana<sup>7</sup>, Karen R Jacobson<sup>8,9</sup>, Thomas C Victor<sup>9</sup>, Robin M Warren<sup>9</sup>, Elizabeth M Streicher<sup>9</sup>, Alistair Calver<sup>10</sup>, Alex Sloutsky<sup>11</sup>, Devinder Kaur<sup>11</sup>, Jamie E Posey<sup>12</sup>, Bonnie Plikaytis<sup>12</sup>, Marco R Oggioni<sup>13</sup>, Jennifer L Gardy<sup>14</sup>, James C Johnston<sup>15</sup>, Mabel Rodrigues<sup>16</sup>, Patrick K C Tang<sup>16</sup>, Midori Kato-Maeda<sup>17</sup>, Mark L Borowsky<sup>18,19</sup>, Bhavana Muddukrishna<sup>18,19</sup>, Barry N Kreiswirth<sup>20</sup>, Natalia Kurepina<sup>20</sup>, James Galagan<sup>2,21–23</sup>, Sebastien Gagneux<sup>24,25</sup>, Bruce Birren<sup>2</sup>, Eric J Rubin<sup>6</sup>, Eric S Lander<sup>2</sup>, Pardis C Sabeti<sup>2–4,6</sup> & Megan Murray<sup>26,27</sup>

*M. tuberculosis* is evolving antibiotic resistance, threatening attempts at tuberculosis epidemic control. Mechanisms of resistance, including genetic changes favored by selection in resistant isolates, are incompletely understood. Using 116 newly sequenced and 7 previously sequenced *M. tuberculosis* whole genomes, we identified genome-wide signatures of positive selection specific to the 47 drug-resistant strains. By searching for convergent evolution—the independent fixation of mutations in the same nucleotide position or gene—we recovered 100% of a set of known resistance markers. We also found evidence of positive selection in an additional 39 genomic regions in resistant isolates. These regions encode components in cell wall biosynthesis, transcriptional regulation and DNA repair pathways. Mutations in these regions could directly confer resistance or compensate for fitness costs associated with resistance. Functional genetic analysis of mutations in one gene, *ponA1*, demonstrated an *in vitro* growth advantage in the presence of the drug rifampicin.

The evolution of antibiotic-resistant bacteria is a major public health concern. To combat antibiotic-resistant infections, not only must new drugs be developed, but existing drugs must also be used more effectively. With some exceptions (for example, in the case of phenotypic drug tolerance), resistance is encoded in the bacterial genome; therefore, resistance-associated mutations, whether they are directly causal in resistance or not, can serve as biomarkers that can be rapidly identified in the clinic by PCR or sequencing-based assays. These molecular biomarkers allow the determination of a bacterial infection's drug resistance profile in a matter of hours, instead

of the days or weeks required for culture-based diagnostics. In some cases, this time lag can make the difference between successful or unsuccessful treatment.

Here we describe a method to identify biomarkers of drug resistance in a rapid and unbiased manner. This method consists of sequencing the genomes of bacteria with different resistance phenotypes and applying phylogenetic methods and statistical tests for positive selection to identify variants in the genome that are consistently associated with resistance. The method is amenable to different microbes with different phenotypes of interest. Here we apply it to identify

<sup>1</sup>Pulmonary and Critical Care Division, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. <sup>2</sup>The Eli and Edythe L. Broad Institute, Cambridge, Massachusetts, USA. <sup>3</sup>Department of Organismic and Evolutionary Biology, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, USA. <sup>4</sup>Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, Massachusetts, USA. <sup>5</sup>Département de Sciences Biologiques, Université de Montréal, Montréal, Quebec, Canada. <sup>6</sup>Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA. <sup>7</sup>Department of Bioinformatics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. <sup>8</sup>Section of Infectious Diseases, Boston University School of Medicine, Boston, Massachusetts, USA. <sup>9</sup>Department of Science and Technology/National Research Foundation Centre of Excellence for Biomedical TB Research, Medical Research Council (MRC) Centre for Molecular and Cellular Biology, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa. <sup>10</sup>Anglogold Ashanti Health West Vaal Hospital, Orkney, South Africa. <sup>11</sup>Massachusetts Supranational TB Reference Laboratory, University of Massachusetts Medical School, Boston, Massachusetts, USA. <sup>12</sup>Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. <sup>13</sup>Department of Genetics, University of Leicester, Leicester, UK. <sup>14</sup>Communicable Disease Prevention and Control Services, British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada. <sup>15</sup>Clinical Prevention Services, British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada. <sup>16</sup>Mycobacteriology/TB Laboratory, Public Health Microbiology and Reference Laboratory, Provincial Health Services Authority Laboratories, British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada. <sup>17</sup>Division of Pulmonary and Critical Care, University of California, San Francisco, San Francisco, California, USA. <sup>18</sup>Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. <sup>19</sup>Department of Genetics, Harvard Medical School, Harvard University, Boston, Massachusetts, USA. <sup>20</sup>Public Health Research Institute Tuberculosis Center, Rutgers, The State University of New Jersey, Newark, New Jersey, USA. <sup>21</sup>Department of Biomedical Engineering, Boston University, Boston, Massachusetts, USA. <sup>22</sup>Department of Microbiology, Boston University, Boston, Massachusetts, USA. <sup>23</sup>Bioinformatics Program, Boston University, Boston, Massachusetts, USA. <sup>24</sup>Swiss Tropical and Public Health Institute, Basel, Switzerland. <sup>25</sup>University of Basel, Basel, Switzerland. <sup>26</sup>Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, USA. <sup>27</sup>Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. <sup>28</sup>These authors contributed equally to this work. Correspondence should be addressed to M.R.F. (mrfarhat@partners.org) or M.M. (mmurray@hsph.harvard.edu).

Received 23 January; accepted 8 August; published online 1 September 2013; doi:10.1038/ng.2747

biomarkers of drug resistance in *M. tuberculosis*, the bacterium responsible for tuberculosis.

The evolution and spread of drug-resistant tuberculosis threaten to undermine the success of tuberculosis treatment and control programs worldwide. Multidrug-resistant (MDR) tuberculosis is defined as tuberculosis that is resistant to isoniazid and rifampicin, the two most effective anti-tubercular drugs. With a global estimate of 650,000 MDR cases in 2010 (ref. 1) and a rising number of cases that are extensively drug resistant (XDR; defined as MDR cases that are also resistant to fluoroquinolones and injectable agents), drug-resistant tuberculosis poses a major challenge, requiring advances in diagnostics, methods of surveillance and therapeutics.

Resistance in *M. tuberculosis* is thought to arise through the serial acquisition of point mutations in genes encoding drug-activating enzymes or drug targets. Current molecular diagnostics amplify and detect known drug resistance-associated mutations, and their performance depends on the inclusion of a comprehensive catalog of these mutations. Although known mutations explain much resistance in *M. tuberculosis*, causative mutations have not been identified in 10–40% of clinically resistant isolates<sup>2</sup>, and, even where causative mutations have been identified, there may be additional variants that contribute to drug resistance.

In addition to classical drug resistance genes (encoding the protein target of the drug or a drug-metabolizing enzyme), mutations in three other classes of genes may confer a selective advantage in the presence of drugs. First, mutations that reduce cell wall permeability or increase the activity of drug efflux pumps are expected to increase the mean inhibitory concentrations of drugs, potentially providing an early step toward full-blown drug resistance<sup>3</sup>. Second, compensatory mutations that ameliorate the fitness costs of other resistance-conferring mutations can occur and may be selected, as seen in studies of both clinical and experimental evolution of drug resistance<sup>4</sup>. Third, mutator phenotypes can increase the rate at which rare beneficial mutations occur (although at the expense of also accumulating deleterious mutations) and therefore provide a selective advantage in the presence of drug treatment<sup>5</sup>.

## RESULTS

To identify genetic markers of drug resistance, we performed next-generation whole-genome sequencing of 116 *M. tuberculosis* isolates from 4 categories: (i) 8 epidemiologically linked clusters of cases (epiclusters) with emergent drug resistance, (ii) 2 uniformly drug-sensitive epiclusters, (iii) 35 non-epidemiologically linked isolates sampled to represent the 6 major global lineages of *M. tuberculosis* and (iv) 8 isolates from a single infected individual displaying emergent resistance (Fig. 1). We combined these data with publicly available genomes from seven isolates. The full sample set of 123 *M. tuberculosis* isolates included 47 isolates resistant to at least 1 tuberculosis drug, including 9 isolates of XDR tuberculosis. The resulting data set captured substantial genetic diversity, with 24,711 polymorphic sites identified relative to the H37Rv reference genome. A genome-wide phylogeny showed substantial population differentiation between epiclusters, which was confirmed by high fixation indices ( $F_{ST} > 0.36$ ) between all epicluster pairs (Fig. 1b,c).

We first determined whether drug resistance could be explained by previously known resistance-conferring mutations. We performed additional deep targeted sequencing of known resistance-associated genes in 35 resistant isolates (including 15 isolates with no apparent resistance-conferring mutations and 20 with at least 1 known resistance-conferring mutation). We detected mutations in known resistance determinants that had been missed by the initial whole-genome

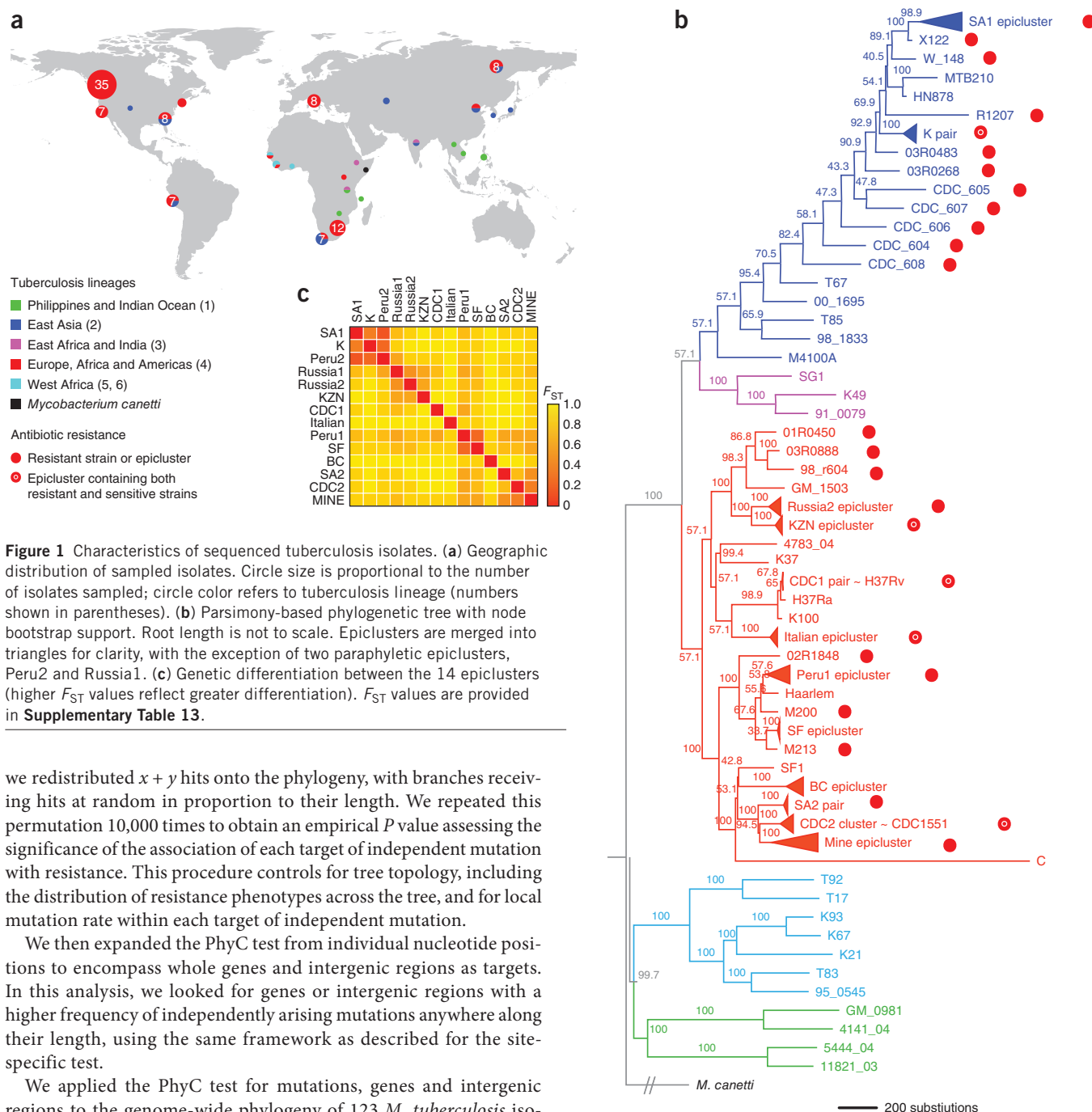
sequencing in 2 isolates; the remaining 13 isolates had confirmed resistance to at least 1 drug that could not be explained by known mutations (Supplementary Table 1).

Next, we reasoned that variants under selection in resistant strains could provide clues about the cellular mechanisms conferring resistance while also serving as biomarkers of resistance. We therefore sought to identify genes harboring mutations that confer a selective advantage to drug-resistant strains. Unfortunately, many commonly used tests to identify genes under positive selection are not well suited to bacteria such as *M. tuberculosis*. Haplotype-based tests for positive selection, often used in humans and other eukaryotes, cannot be used, as genetic diversity in *M. tuberculosis* arises primarily by clonal expansion rather than by mating and homologous recombination among isolates<sup>6,7</sup>. The widely used dN/dS method is also not suited to *M. tuberculosis*, as it has low sensitivity in detecting positive selection in recently diverged sequences from a single species<sup>8</sup> and as strong purifying selection on synonymous mutations in *M. tuberculosis*<sup>6</sup> can spuriously give rise to high dN/dS scores, resulting in low specificity. Indeed, the dN/dS method lacked power and, likely, specificity when applied to our *M. tuberculosis* data set, recovering only 5 of 11 known resistance determinants, while detecting 143 additional genes (Supplementary Table 2).

Instead, we sought to leverage evolutionary convergence—the repeated and independent emergence of resistance-associated mutations at specific loci or genes—to develop a test for selection in clonal bacterial species such as *M. tuberculosis*<sup>7</sup>. To identify independently arising mutations, we reconstructed a phylogenetic tree for the 123 isolates using *Mycobacterium canettii* as an outgroup. On the basis of this tree, we inferred nonsynonymous and noncoding ancestral sequence changes and internal resistance states using parsimony. We focused on nonsynonymous mutations, as these were more likely to encode functional protein changes than their synonymous counterparts. Nevertheless, because of emerging evidence that synonymous sites may also be under selection for adaptive changes in gene expression or mRNA stability<sup>6,9</sup>, we also inferred synonymous mutations (in a secondary analysis reported in the Supplementary Note).

We took several precautions to ensure that the reconstructed genetic changes and resistance states in our analysis were not influenced by possible errors in tree topology. First, we reconstructed the phylogeny in triplicate using different methodologies and removed all mutations not inferred in all three trees. We also ignored ambiguous mutations from the ancestral reconstruction and mutations occurring at branches with lower than 70% bootstrap support. Second, to remove local uncertainty in tree topology, we counted ‘close changes’ only once. Close changes were defined as any changes that occurred in two isolates separated by less than the 98th percentile for within-epicluster genetic distance. Third, we implemented a simplified pairwise convergence test in which we compared the most sensitive isolate to the most resistant isolate in each of eight epiclusters, ignoring the rest of the phylogeny entirely (Supplementary Fig. 1).

Using the resulting high-confidence ancestral reconstruction, we designed a phylogenetic convergence test (phyC). We first looked for specific mutations with a higher frequency in the resistant branches compared with in the sensitive branches as candidate targets of independent mutation. To distinguish convergence due to positive selection in resistant branches from patterns expected by chance under neutral evolution<sup>10</sup>, we assessed the significance of the difference for each candidate target of independent mutation in its distribution relative to the distribution expected based on observed mutations across the phylogeny. Briefly, for each target of independent mutation with mutations in  $x$  resistant branches and  $y$  sensitive branches,



we redistributed  $x + y$  hits onto the phylogeny, with branches receiving hits at random in proportion to their length. We repeated this permutation 10,000 times to obtain an empirical  $P$  value assessing the significance of the association of each target of independent mutation with resistance. This procedure controls for tree topology, including the distribution of resistance phenotypes across the tree, and for local mutation rate within each target of independent mutation.

We then expanded the PhyC test from individual nucleotide positions to encompass whole genes and intergenic regions as targets. In this analysis, we looked for genes or intergenic regions with a higher frequency of independently arising mutations anywhere along their length, using the same framework as described for the site-specific test.

We applied the PhyC test for mutations, genes and intergenic regions to the genome-wide phylogeny of 123 *M. tuberculosis* isolates. For our analysis, we defined the resistance phenotype as resistance to any tuberculosis drug as determined by conventional drug susceptibility testing so that we would be able to identify mutations associated with multiple resistance phenotypes as well as those that conferred resistance to a single drug. We repeated these analyses to identify selection in isolates resistant to each of the five first-line tuberculosis drugs (isoniazid, rifampin, pyrazinamide, ethambutol and streptomycin).

As a proof of concept, we assessed the functional impact of the observed mutations within one of the targets of independent mutation identified by the PhyC test. We constructed two *ponA1* mutants (carrying two of the three SNPs (c.123C>G and c.1095G>T) that were most enriched in resistant strains) in an H37Rv *M. tuberculosis* laboratory strain using recombinering and site-directed mutagenesis.

We then compared the survival of the two mutant strains to that of wild-type cells and cells that lacked the *ponA1* gene under increasing concentrations of the drugs rifampicin, isoniazid, streptomycin and ofloxacin.

### Targets of independent mutation

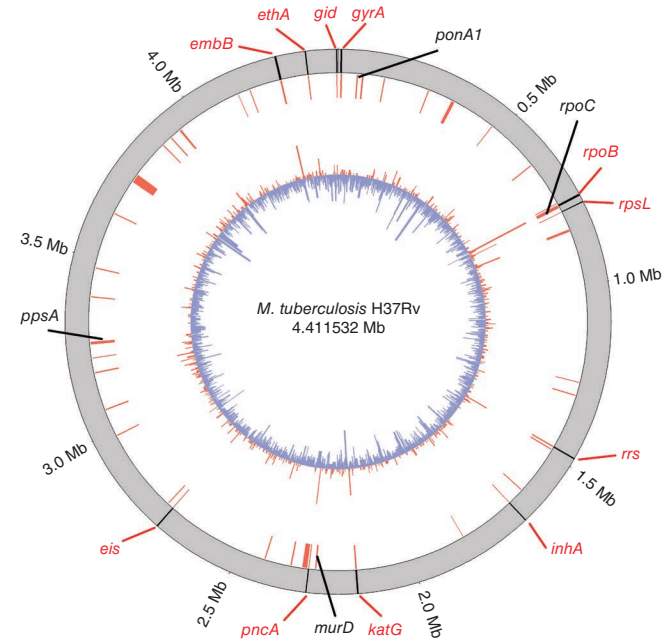
PhyC detected all 11 known resistance determinants as significant targets of independent mutation. Nine of these determinants were also identified by a weaker but conservative phylogeny-independent pairwise convergence test (Supplementary Tables 3 and 4) developed here. We further identified 39 new targets of independent mutation not previously associated with resistance, consisting of 7 nonsynonymous coding sites, 2 noncoding sites, 28 genes and 2

**Figure 2** Candidate genes under selection in resistant *M. tuberculosis*. Circular plot of gene locations. Outer black lines represent the 11 benchmark drug resistance-associated genes in the H37Rv reference genome (red text). Inner red lines represent the locations of targets of independent mutation. Four new targets of independent mutation of interest are shown in black text. The innermost bar plot shows the number of mutations per gene or intergenic region in resistant (red) and sensitive (blue) isolates. Plotted using Circos<sup>29</sup>.

intergenic regions ( $P < 0.05$ ) (Fig. 2 and Supplementary Table 5). All nine single-nucleotide targets of independent mutation fell within genes or intergenic regions also identified as targets of independent mutation. We observed that mutations in resistant branches clustered more closely in the genome than those in sensitive branches and that many of the targets of independent mutation fell in these regions of dense, resistance-specific mutations (Supplementary Tables 6 and 7). The 946T nucleotide allele in the gene encoding the conserved membrane protein Rv0218 had the highest number of independent hits in resistant branches of any candidate mutation, occurring in eight resistant branches and on no sensitive branch ( $P < 0.00001$ ). However, because there were more sensitive than resistant isolates in our data set, mutations in sensitive branches are far more prevalent than in resistant branches (compare the red and blue histograms in Fig. 2). Several of the targets of independent mutation significantly associated with resistance also occurred in sensitive branches (Supplementary Table 5). This finding suggests that several targets of independent mutation may not cause resistance directly but may rather provide an incremental fitness advantage to resistant strains.

Functions of candidate selected loci

Of the 39 newly associated genomic regions identified by PhyC, 11 have annotated functions (Table 1), 16 belong to a family of genes (PE/PPE) of principally unknown function that is unique to mycobacteria and constitutes about 10% of the *M. tuberculosis* genome (Supplementary Table 5) and the remaining 12 are of unknown function. We systematically mined the literature for the genes not previously associated with resistance, noting evidence that many closely interact with known drug resistance genes (physically or genetically) or drug efflux pumps, alter intrinsic drug resistance in *M. tuberculosis* or non-tuberculous mycobacteria, are involved in DNA repair, replication or recombination, or affect cell wall biogenesis (Supplementary Fig. 2).



Previously known resistance loci

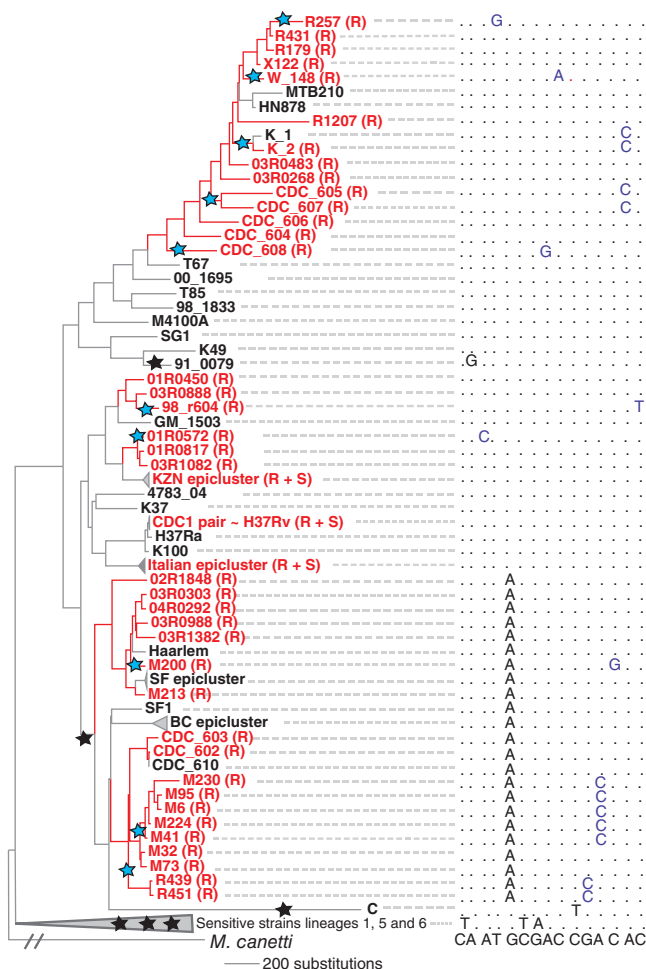
Two new targets of independent mutation were located close to known resistance-associated genes in the genome, suggesting that they modify or compensate for the phenotypes of mutations in the known genes. The first occurred in the promoter of the known resistance-associated gene *rrs*, which encodes the 16S RNA component of the ribosome, a target of aminoglycoside drugs. The second, *rpoC* (Fig. 3), is in the same operon as *rpoB*, which codes for the  $\beta$  subunit of RNA polymerase, the main target of the drug rifampicin. *rpoC* encodes the interacting  $\beta'$  subunit and has been identified as a target of compensatory mutations modifying the fitness of rifampicin-resistant isolates of both *M. tuberculosis* and *Salmonella typhimurium*<sup>11–14</sup>. Previous studies have shown that substitutions in *rpoC* are frequent in clinical rifampicin-resistant isolates with known *rpoB* mutations and that the relative fitness of *rpoC* and *rpoB* double mutants *in vitro* is higher than for strains with *rpoB* mutations alone<sup>12,13</sup>. Some *rpoC* mutations in *S. typhimurium* also confer low-level rifampicin resistance, suggesting that rifampin resistance phenotypes may be the result

**Table 1** Targets of independent mutation with annotated function

Gene	Rv number	Cellular function					Resistance association		
		Synthesis or regulation of surface-exposed lipids	Peptidoglycan homeostasis	Transcriptional regulation	DNA replication and repair	Glucose metabolism and antioxidant	Associated with resistance in <i>M. tuberculosis</i>	Associated with resistance in NTM	Associated with resistance in other bacteria
<i>ppsA</i>	<i>Rv2931</i>	19					27	25	
<i>pks3</i>	<i>Rv1180</i>	21							
<i>pks12</i>	<i>Rv2048c</i>	20					24	24,26	
<i>ponA1</i>	<i>Rv0050</i>		22					23	30
<i>murD</i>	<i>Rv2155c</i>		22						
<i>mtrB</i>	<i>Rv3245c</i>			31				32,33	34
<i>rpoC</i>	<i>Rv0668</i>				12		11,12		13
<i>dnaQ</i>	<i>Rv3711c</i>				35				15
<i>opcA</i>	<i>Rv1446c</i>					35	36		
<i>rbsK</i>	<i>Rv2436</i>				37	35			
<i>rrs</i> promoter (pre- <i>Rvn01</i> )							38	39	

Numbers refer to literature references. Genes involved in cell wall biosynthesis include *ppsA*, *pks3*, *pks12*, *ponA1* and *murD*. A complete list of targets of independent mutation is given in Supplementary Table 5. NTM, non-tuberculous mycobacteria.





**Figure 3** Evolutionary convergence at the gene level in *rpoC*. Resistant branches (inferred by parsimony and usually involving progressive resistance to several drugs) and strain names are shown in red; sensitive branches are shown in black. Stars on the phylogeny designate inferred sequence changes in *rpoC*: blue, changes in resistant branches (ten in total); black, changes in sensitive branches (six in total). Nucleotides in the multiple-sequence alignment are also colored blue or black accordingly. Sites shown in the multiple-sequence alignment (left to right) are 763,884, 764,181, 764,580, 764,819, 765,150, 765,171, 765,230, 765,462, 765,463, 765,482, 765,619, 766,467, 766,488, 766,645 and 767,060 (H37Rv coordinates).

result in strong mutator phenotypes<sup>15</sup>. So far, no similar phenotype has been described in *M. tuberculosis*, although *dnaQ* variants are not uncommon among clinical isolates<sup>16</sup>.

### The PE/PPE gene family as a target of independent mutation

Sixteen new targets of independent mutation are members of the PE/PPE family, with the majority within the PGRS subfamily; this was the only gene group significantly enriched for in the convergence analysis. Multiple members of this family encode surface-exposed cell wall proteins; some affect cell wall structure and permeability, and some have been shown to be antigens<sup>17</sup>. PE/PPE genes contain an extremely high density of substitutions and were therefore excluded from the genome-wide phylogeny. As a result, it was expected (although not guaranteed) that these genes would be enriched for conflicting phylogenetic signal (homoplasy), although homoplastic mutations may not necessarily be associated with resistance. The association of PE/PPE genes with drug resistance is therefore noteworthy. Owing to the high genetic diversity in these genes, associations may reflect random fixation of this diversity during population bottlenecks that occur during antibiotic treatment rather than genuine positive selection on resistant isolates. In other words, resistant isolates may be descended from the survivors of severe bottlenecks during which neutral mutations repeatedly become fixed, and these mutations are most readily observed in diverse loci such as PE/PPE genes. However, we cannot rule out a possible functional role for PE/PPE genes in the evolution of resistance—for example, *PPE60* is one of the best candidates to account for some of the unexplained kanamycin-resistant isolates (see below and **Supplementary Table 8**).

of additive substitutions in different genes<sup>13</sup>. Of the 43 rifampicin-resistant isolates in our collection, 13 (30%) harbored *rpoC* substitutions that did not appear in any rifampicin-sensitive isolates.

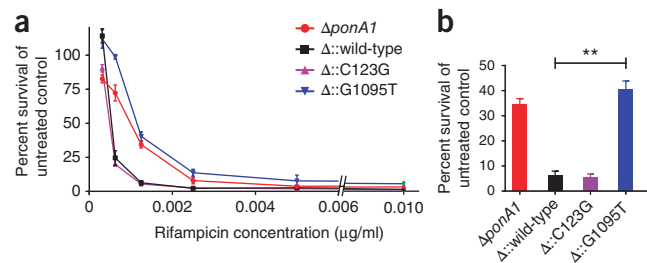
We determined whether nonsynonymous mutations in the targets of independent mutation could explain resistance in the 13 isolates in which we did not find any known drug resistance-conferring mutations. For each drug and isolate, we identified all mutations in the candidate genes, excluding mutations occurring in any isolate sensitive to that drug. Although no single candidate mutation or gene was found to account for all the unexplained drug-specific resistance, two of six isolates with unexplained kanamycin resistance harbored mutations in *PPE60* (**Supplementary Table 8**). Eight of the 13 isolates (62%) exhibited changes in at least 1 target of independent mutation, with 4 of the isolates exhibiting changes in 2 or more.

### Drug efflux pumps

Although no efflux pumps were identified among genes that met our statistical criteria for targets of independent mutation, we found that several efflux pumps, including the ABC transporters Rv0194 and Rv1463, were affected by a larger number of independent mutations in resistant strains relative to sensitive strains.

### DNA repair

Another of the new targets of independent mutation, *dnaQ*, encodes a component of DNA polymerase III that provides proofreading activity during DNA replication. Several *dnaQ* mutations in *Escherichia coli*



**Figure 4** *M. tuberculosis ponA1* mutant survival in the presence of the drug rifampicin. **(a)** Bacterial survival (percent of untreated control OD<sub>600</sub> absorbance) under increasing concentrations of rifampicin for  $\Delta$ *ponA1* (*ponA1* deletion mutant),  $\Delta$ ::wild-type (*ponA1* deletion mutant complemented with the wild-type *ponA1* gene),  $\Delta$ ::G1095T (*ponA1* deletion mutant complemented with *ponA1* carrying the c.1095G>T allele) and  $\Delta$ ::C123G (*ponA1* deletion mutant complemented with *ponA1* carrying the c.123C>G allele). **(b)** Bacterial survival as in **a** of strains cultured in the presence of 0.00125  $\mu$ g/ml rifampicin. Two-sided *t* tests were used to evaluate significance. The difference in survival between the  $\Delta$ ::wild-type and  $\Delta$ ::G1095T strains was significant (\*\*), with a *P* value of 0.006. Error bars, s.d. Four replicate experiments were performed with mean values shown.

## Cell wall homeostasis

Five new targets of independent mutation contribute to *M. tuberculosis* cell wall biogenesis or remodeling. The structure of the mycobacterial cell wall is unique among prokaryotes in that, in addition to the peptidoglycan layer typical of most bacteria, it contains several outer layers characterized by unusual, complex lipids (Table 1 and Supplementary Fig. 3). These layers contribute to the permeability barrier that underlies the intrinsic antibiotic resistance of most mycobacteria<sup>18</sup>. Multiple tuberculosis drugs target structures in the cell wall, and many of the known resistance-associated genes code for enzymes in cell wall lipid pathways. Three of the five genes (*ppsA*, *pks12* and *pks3*) participate in the biosynthesis and translocation of surface-exposed lipids, including phthiocerol dimycocerosate (PDIM)<sup>19–21</sup>, whereas the remaining two (*murD* and *ponA1*) contribute to the biosynthesis and homeostasis of the cell wall component peptidoglycan<sup>22</sup>. Deletion of *ppsA* and depletion of *pks12* or *ponA1* each affect susceptibility to antibiotics in non-tuberculous mycobacteria and/or *M. tuberculosis*<sup>23–25</sup>. Deletion of *pks12* has recently been shown to increase drug resistance in *Mycobacterium avium* through a cell wall–remodeling pathway<sup>26</sup>. In addition, *pks12* had a synonymous site that was a significant target of independent mutation (Supplementary Table 9).

## Functional analysis of PonA1 mutants

In the presence of the drug rifampicin, the strain carrying the *ponA1* c.1095G>T mutation had a 4- to 6-fold survival advantage over wild-type strains at a rifampicin concentration of 0.00125 µg/ml (Fig. 4). The minimal inhibitory concentration (MIC) of rifampicin for this mutant was estimated at 0.0025 µg/ml, twofold higher than the wild-type MIC (0.00125 µg/ml). In contrast, the mutant strain encoding *ponA1* c.123C>G showed no growth advantage in the presence of rifampicin, and neither mutant demonstrated a growth advantage when grown in the presence of isoniazid, streptomycin or ofloxacin. The c.1095 nucleotide site maps close to the PonA1 transpeptidase domain catalytic site, raising the possibility that the SNP inactivates enzymatic activity; this idea is supported by the finding that the *ponA1* deletion mutant had a similar rifampicin resistance phenotype (Fig. 4).

## DISCUSSION

This work describes a comprehensive genome-wide screen for genes under selection in clinical drug-resistant *M. tuberculosis* isolates. Our method involves sequencing the genomes of related bacteria with different phenotypes of interest (in this case, antibiotic resistance), reconstructing a phylogeny and identifying targets of convergent evolution using a simple statistical test. This selection test is highly sensitive and was demonstrated to recover all of a set of 11 known drug resistance markers. The method is also generalizable. Although the method relies on a genome-wide phylogeny, it can accommodate recombination, provided that it is not so rampant as to completely obscure the phylogeny. Recombinant regions often conflict with the genome-wide phylogeny, allowing them to be identified by our method as targets of convergent evolution if they are consistently associated with the phenotype of interest but not if they are randomly distributed among genomes with different phenotypes. The method is therefore amenable to bacteria with a range of recombination rates. It provides a rapid and unbiased means of identifying molecular biomarkers that are predictive of phenotypes of interest.

Applying this method to *M. tuberculosis*, we identified 39 genes and intergenic regions newly associated with resistance. Although several of the selected mutations occur in genes that are either close

to loci known to be associated with drug resistance or are mutator genes in other organisms, the preponderance of regions were associated with cell wall permeability phenotypes. This finding suggests that stable drug resistance phenotypes may evolve through a complex stepwise process involving cell wall remodeling. Our finding that a *ponA1* mutation (c.1095G>T) identified as a target of independent mutation conferred a fitness advantage in the presence of rifampicin is consistent with this model.

The relevance of cell wall–remodeling pathways to drug resistance is also highlighted by two recent studies that compared resistant isolates to their drug-sensitive precursors. In one study, the investigators noted increased levels of PDIM and peptidoglycan precursors and upregulation of the PDIM biosynthetic operon (including *ppsA*) in rifampicin-resistant strains. Interestingly, we identified *ppsA* as a target of independent mutation in strains resistant to rifampicin (95% of which had nonsynonymous mutations in *rpoB*; Supplementary Table 10) in addition to other drugs, raising the possibility that rifampin resistance–causing *rpoB* mutations may lead to alterations in cell wall metabolism, possibly as a result of altered transcription<sup>27</sup>. The second study documented the occurrence of 11 new nonsynonymous mutations in serial *M. tuberculosis* isolates from 3 infected individuals who developed increasing levels of resistance during treatment for tuberculosis. Seven of the mutations occurred in genes involved in cell wall biosynthesis or transport, including in *fadD32* and *Rv1739c*, which encodes an ABC transporter. Although none of these genes overlapped with the targets of independent mutation identified in this study, the enrichment of mutations in genes associated with cell wall biosynthetic pathways among progressively drug-resistant strains is consistent with our findings and further supports the hypothesis that these changes reflect accommodation to drug exposure<sup>28</sup>.

The genomic targets of independent mutation we have identified here are associated with drug resistance and may represent changes that confer a selective advantage in the presence of drug. In aggregate, they provide promising new targets for molecular diagnostics and the development of new therapeutics for a range of drug resistance phenotypes in *M. tuberculosis*.

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

**Accession codes.** All sequences have been rendered publicly available through NCBI. The Haarlem, C, 98\_r168 and w-148 genome assemblies are available under GenBank accessions [AASN000000000](#), [AAKR000000000](#), [ABVM000000000](#) and [ACSX000000000](#), respectively. Raw sequences for the 35 isolates from Vancouver are available at the NCBI Sequence Read Archive (SRA) under accession [SRA020129](#). KZN-DS (KZN-4207), KZN\_MDR (KZN-1435) and KZN\_XDR (KZN-605) raw sequence reads are available under accession [SRA009637](#). Raw sequence data for the isolates corresponding to our study identification numbers 51–73 are available under accession [SRA009341](#). Read sequences for the rest of our isolates are available under accession [SRA009458](#) with the project name XDR comparative. Publicly available partial or complete genome sequences for MTB210, H37Ra, HN878, R1207 and X122 were accessed from GenBank accessions [ADAB000000000](#), [AAYK010000000](#), [ADNF010000000](#), [ADNH010000000](#) and [ADNG010000000](#), respectively.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We thank the technical staff of the British Columbia Centre for Disease Control Public Health Microbiology and Reference Mycobacteriology Laboratory in Vancouver, M. Bosman from the National Health Laboratory Service in Cape Town and L. Fattorini from the Istituto Superiore di Sanita in Rome. This work was funded by a Senior Ellison Foundation Award (M.M.) and in part by a contact from the National Institute of Allergy and Infectious Diseases (HHSN266200400001C to B.B.), the Department of Pulmonary and Critical Care at Massachusetts General Hospital (M.R.F.), a postdoctoral fellowship from the Harvard MIDAS Center for Communicable Disease Dynamics (B.J.S.) and a Packard Foundation Fellowship (P.C.S.). S.G. was supported by the Swiss National Science Foundation (PP0033\_119205).

## AUTHOR CONTRIBUTIONS

This study was designed and conducted by M.R.F. and M.M. M.R.F. wrote the first drafts of the manuscript. B.J.S., P.C.S. and E.S.L. provided conceptual input on the evolutionary testing, analysis support and key manuscript edits. K.J.K. and E.J.R. constructed the *ponA1* mutants and measured their MICs. R.S. provided bioinformatics support and K.R.J. helped with the curation of the isolate phenotypes. R.M.W., E.M.S., T.C.V. and A.C. conducted molecular epidemiological studies and performed molecular characterization, drug susceptibility testing and selection of isolates from South Africa. A.S. and D.K. performed molecular characterization and drug sensitivity testing and selected isolates from Peru and Russia. B.P. and J.E.P. performed molecular characterization, drug sensitivity testing and selection of isolates from the Centers for Disease Control and Prevention. M.R.O. identified the individual with progressively resistant tuberculosis and performed molecular characterization and selection of serial isolates from this individual in Italy. J.L.G., J.C.J., M.R. and P.K.C.T. conducted the tuberculosis outbreak investigation in British Columbia and performed molecular characterization, drug susceptibility testing and sequencing of these isolates. M.K.-M. conducted the epidemiological study of tuberculosis transmission in San Francisco, and M.L.B. and B.M. performed molecular characterization and sequencing of these isolates. B.N.K. and N.K. characterized the W-148, Haarlem and C isolates. S.G. collected the 24 drug-sensitive *M. tuberculosis* diversity strain set. J.G. and B.B. provided oversight for sequencing and bioinformatics support.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- World Health Organization. *Global Tuberculosis Control 2011* (World Health Organization Press, Geneva, 2011).
- Campbell, P.J. *et al.* Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **55**, 2032–2041 (2011).
- Nikaido, H. Prevention of drug access to bacterial targets: permeability barriers and active efflux. *Science* **264**, 382–388 (1994).
- Schrag, S.J., Perrot, V. & Levin, B.R. Adaptation to the fitness costs of antibiotic resistance in *Escherichia coli*. *Proc. Biol. Sci.* **264**, 1287–1291 (1997).
- Denamur, E. & Matic, I. Evolution of mutation rates in bacteria. *Mol. Microbiol.* **60**, 820–827 (2006).
- Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E.P.C. After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.* **22**, 721–734 (2012).
- Shapiro, B.J., David, L.A., Friedman, J. & Alm, E.J. Looking for Darwin's footprints in the microbial world. *Trends Microbiol.* **17**, 196–204 (2009).
- Kryazhimskiy, S. & Plotkin, J.B. The population genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).
- Agashe, D., Martinez-Gomez, N.C., Drummond, D.A. & Marx, C.J. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol. Evol.* **30**, 549–560 (2013).
- Rokas, A. & Carroll, S.B. Frequent and widespread parallel evolution of protein sequences. *Mol. Biol. Evol.* **25**, 1943–1953 (2008).
- Casali, N. *et al.* Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.* **22**, 735–745 (2012).
- Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106–110 (2012).
- Brandis, G., Wrande, M., Liljas, L. & Hughes, D. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. *Mol. Microbiol.* **85**, 142–151 (2012).
- de Vos, M. *et al.* Putative compensatory mutations in the *rpoC* gene of rifampin-resistant *Mycobacterium tuberculosis* are associated with ongoing transmission. *Antimicrob. Agents Chemother.* **57**, 827–832 (2013).
- Tanabe, K., Kondo, T., Onodera, Y. & Furusawa, M. A conspicuous adaptability to antibiotics in the *Escherichia coli* mutator strain, dnaQ49. *FEMS Microbiol. Lett.* **176**, 191–196 (1999).
- Dos Vultos, T., Mestre, O., Tonjum, T. & Gicquel, B. DNA repair in *Mycobacterium tuberculosis* revisited. *FEMS Microbiol. Rev.* **33**, 471–487 (2009).
- Soldini, S. *et al.* PPE\_MPTR genes are differentially expressed by *Mycobacterium tuberculosis* in vivo. *Tuberculosis (Edinb.)* **91**, 563–568 (2011).
- Kaur, D., Guerin, M.E., Skovierová, H., Brennan, P.J. & Jackson, M. Chapter 2: biogenesis of the cell wall and other glycoconjugates of *Mycobacterium tuberculosis*. *Adv. Appl. Microbiol.* **69**, 23–78 (2009).
- Yu, J. *et al.* Both phthiocerol dimycocerosates and phenolic glycolipids are required for virulence of *Mycobacterium marinum*. *Infect. Immun.* **80**, 1381–1389 (2012).
- Matsunaga, I. *et al.* *Mycobacterium tuberculosis pks12* produces a novel polyketide presented by CD1c to T cells. *J. Exp. Med.* **200**, 1559–1569 (2004).
- Dubey, V.S., Sirakova, T.D. & Kolattukudy, P.E. Disruption of *msl3* abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in *Mycobacterium tuberculosis* H37Rv and causes cell aggregation. *Mol. Microbiol.* **45**, 1451–1459 (2002).
- Hett, E.C., Chao, M.C. & Rubin, E.J. Interaction and modulation of two antagonistic cell wall enzymes of mycobacteria. *PLoS Pathog.* **6**, e1001020 (2010).
- Billman-Jacobe, H., Haites, R.E. & Coppel, R.L. Characterization of a *Mycobacterium smegmatis* mutant lacking penicillin binding protein 1. *Antimicrob. Agents Chemother.* **43**, 3011–3013 (1999).
- Philalay, J.S., Palermo, C.O., Hauge, K.A., Rustad, T.R. & Cangelosi, G.A. Genes required for intrinsic multidrug resistance in *Mycobacterium avium*. *Antimicrob. Agents Chemother.* **48**, 3412–3418 (2004).
- Chavadi, S.S. *et al.* Inactivation of *tesA* reduces cell wall lipid production and increases drug susceptibility in mycobacteria. *J. Biol. Chem.* **286**, 24616–24625 (2011).
- Matsunaga, I., Meda, S., Nakata, N. & Fujiwara, N. The polyketide synthase-associated multidrug tolerance in *Mycobacterium intracellulare* clinical isolates. *Chemotherapy* **58**, 341–348 (2012).
- Bisson, G.P. *et al.* Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, *rpoB* mutant *Mycobacterium tuberculosis*. *J. Bacteriol.* **194**, 6441–6452 (2012).
- Sun, G. *et al.* Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J. Infect. Dis.* **206**, 1724–1733 (2012).
- Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
- Shigemura, K. *et al.* Presence of a mutation in *ponA1* of *Neisseria gonorrhoeae* in numerous clinical samples resistant to various  $\beta$ -lactams and other, structurally unrelated, antimicrobials. *J. Infect. Chemother.* **11**, 226–230 (2005).
- Zahrt, T.C. & Deretic, V. An essential two-component signal transduction system in *Mycobacterium tuberculosis*. *J. Bacteriol.* **182**, 3832–3838 (2000).
- Nguyen, H.T., Wolff, K.A., Cartabuke, R.H., Ogowang, S. & Nguyen, L. A lipoprotein modulates activity of the MtrAB two-component system to provide intrinsic multidrug resistance, cytokinetic control and cell wall homeostasis in *Mycobacterium*. *Mol. Microbiol.* **76**, 348–364 (2010).
- Cangelosi, G.A. *et al.* The two-component regulatory system *mtrAB* is required for morphotypic multidrug resistance in *Mycobacterium avium*. *Antimicrob. Agents Chemother.* **50**, 461–468 (2006).
- Möker, N. *et al.* Deletion of the genes encoding the MtrA-MtrB two-component system of *Corynebacterium glutamicum* has a strong influence on cell morphology, antibiotics susceptibility and expression of genes involved in osmoprotection. *Mol. Microbiol.* **54**, 420–438 (2004).
- Lew, J.M., Kapopoulou, A., Jones, L.M. & Cole, S.T. TubercuList—10 years after. *Tuberculosis (Edinb.)* **91**, 1–7 (2011).
- Jiang, X. *et al.* Comparison of the proteome of isoniazid-resistant and -susceptible strains of *Mycobacterium tuberculosis*. *Microb. Drug Resist.* **12**, 231–238 (2006).
- Yang, Q., Liu, Y., Huang, F. & He, Z.-G. Physical and functional interaction between D-ribokinase and topoisomerase I has opposite effects on their respective activity in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *Arch. Biochem. Biophys.* **512**, 135–142 (2011).
- Sandgren, A. *et al.* Tuberculosis drug resistance mutation database. *PLoS Med.* **6**, e2 (2009).
- Nessar, R., Reyart, J.M., Murray, A. & Gicquel, B. Genetic analysis of new 16S rRNA mutations conferring aminoglycoside resistance in *Mycobacterium abscessus*. *J. Antimicrob. Chemother.* **66**, 1719–1724 (2011).



## ONLINE METHODS

**Institutional review boards.** The study was evaluated by institutional review boards at the Harvard School of Public Health, the Broad Institute of MIT and Harvard and the Centers for Disease Control and Prevention, where it was determined that it met criteria for exemption from human subject review.

**Isolate selection.** We assembled an archive of sensitive and resistant isolates that capture a range of *M. tuberculosis* lineages, geographic sources and resistance profiles. Building on our previous molecular epidemiological studies (Supplementary Table 11 and references therein), we aimed to include sets of progressively resistant isolates, sampled either from community transmission chains or from individuals with chronic disease. These sets included isolates from 11 such microepidemics (epicenter isolates), as defined by molecular fingerprinting, chosen to include the most sensitive and most resistant members of the cluster as well as 8 progressively resistant isolates from a single infected individual collected over time. To obtain a measure of background evolution restricted to sensitive isolates and avoid misidentifying highly variable loci as associated with drug resistance, we included 23 geographically diverse drug-sensitive isolates and additional isolates from 2 drug-sensitive epicenters.

To increase the diversity of the sample, we included 11 additional resistant isolates, even when a less resistant progenitor was not available. We also included three isolates that had been spontaneously evolved *in vitro* and manifested an aminoglycoside resistance phenotype that was unexplained by targeted sequencing of all previously known resistance genes. In all, 116 *M. tuberculosis* isolates that were selected for sequencing are described in detail in Supplementary Table 11a.

We added seven publicly available *M. tuberculosis* genomes to our alignment, including two drug-sensitive Beijing-lineage isolates, which were missing from our sampled isolates thus far. Isolates obtained from public sources are detailed in Supplementary Table 11b. There is no established method to determine the power of genomic analyses performed with this sample size, but the strain set studied here is among the largest set of drug-resistant *M. tuberculosis* strains sequenced so far.

Of the 123 total isolates, 47 were resistant to 1 or more drug (Supplementary Table 11). Eighty-three isolates belonged to 14 distinct epicenters, and the rest were isolates with a unique molecular fingerprint. Twelve clusters had one or more resistant isolates. One other publicly available genome from the species *M. canettii* was included to serve as a phylogenetic outgroup.

**Resistance phenotype.** We defined a 'broad resistance' phenotype as resistance to any tuberculosis drug as determined by conventional drug susceptibility testing (Supplementary Table 12). We used this broad resistance as our primary phenotype of interest, as our goal was to identify genes and mutations associated with resistance to at least one drug but potentially to many. As a secondary, more specific phenotype of interest, we used resistance to each of the five first-line tuberculosis drugs (isoniazid, rifampin, pyrazinamide, ethambutol and streptomycin). Resistance status to first-line tuberculosis drugs had much fewer missing data points than resistance to the other drugs (Supplementary Table 11).

**Sequencing, alignment and SNP calling.** DNA was extracted from all isolates using standard methods and was sequenced on an Illumina Genome Analyzer IIx instrument using reads of 36 bp in length or more. Sequence reads were aligned to the reference genome sequence for H37Rv using Mapping and Assembly with Qualities (MAQ)<sup>40</sup>. Reads that aligned with more than three mismatches in the first 24 bp or that aligned to multiple locations were discarded. SNPs were called with a minimum depth of 20× and a consensus quality score of 20 (Supplementary Figs. 4 and 5). The required maximum mapping quality of reads covering each SNP was set at 50 (Supplementary Fig. 6). SNPs within 5 bp of an indel (insertion or deletion) or that did not have an adjacent consensus quality score of 20 were also discarded. Further details are given in the Supplementary Note.

**$F_{ST}$  and dN/dS analyses.** Fixation indices ( $F_{ST}$  values) and dN/dS rates were computed using standard methods detailed in Supplementary Table 13 and the Supplementary Note.

**Phylogeny construction.** The phylogeny was constructed on the basis of multiple-sequence alignment of the *M. tuberculosis* whole-genome sequences, as *M. tuberculosis* populations are thought to be predominantly clonal, with most of the genome supporting a single consensus phylogeny not affected substantially by recombination<sup>6</sup>. A superset of SNPs relative to reference strain H37Rv<sup>35</sup> was created across all clinical isolates from the MAQ SNP reports. SNPs occurring in repetitive elements, including transposases, PE/PPE and PGRS genes, and phiRV1 members (273 genes; 10% of the genome; genes listed in ref. 41) were excluded to avoid any concern about inaccuracies in read alignment in those portions of the genome. Furthermore, SNPs in an additional 39 genes previously associated with drug resistance<sup>38</sup> were also removed to exclude the possibility that homoplasy of drug resistance-associated mutations would substantially alter the phylogeny. After these filters were applied to the initial set of 24,711 SNPs, the 23,393 remaining SNPs were concatenated and used to construct phylogenetic trees with three methods. Using PHYLIP dnaphars algorithm v3.68 (ref. 42), we constructed a parsimony-based phylogenetic tree using default parameters with *M. canettii* as an outgroup root. We constructed a second phylogeny with Bayesian Markov chain Monte Carlo (MCMC) methods as implemented in the package MrBayes v3.2 (ref. 43) using the GTR model and a stop criterion of a standard deviation of split frequencies of <0.05. We constructed a maximum-likelihood tree using PhyML v3.0 (ref. 44) using the GTR model with eight categories for the gamma model with and without *M. canettii* to determine the location of the root. One hundred bootstrap resamplings were performed for each tree, except for the Bayesian tree, where posterior probabilities on the branches were used as a measure of confidence. A phylogeny was also constructed using the full SNP set (without excluding SNPs in repetitive elements or known drug resistance-associated genes), with only minor differences in the terminal branches of the tree found. We used the trees constructed with the exclusion of SNPs in repetitive elements and known drug resistant-associated genes for all subsequent analyses.

**Phylogenetic convergence test for selection (PhyC).** The phylogenetic convergence test used sequences from all branches of the phylogeny. Ancestral nonsynonymous (or intergenic) nucleotide substitutions were reconstructed along each branch using both parsimony and maximum-likelihood criteria in the R v2.14.1 package ape v3.0.1 (ref. 45). Each branch was assigned a 'resistant' or 'sensitive' label using parsimony. Reconstruction was performed in triplicate for the three phylogenies (Bayesian, parsimony and maximum likelihood). We excluded all ambiguously reconstructed states (<90% probability for maximum-likelihood reconstruction). We considered changes occurring along the terminal and deep branches of the phylogenetic tree but excluded changes occurring at branches with bootstrap support or posterior probability of <70%. For each nucleotide position in the genome, we counted the number of convergent SNPs (changes to the same base) in resistant and sensitive branches. Given that some background convergence is expected owing to neutral mutation and sequence error, even without positive selection<sup>10</sup>, we assessed the significance of each convergent SNP compared to the empirical background distribution (Supplementary Fig. 7). For a SNP that converges in  $x$  resistant and  $y$  sensitive branches, we sampled  $x + y$  branches from the distribution of all SNPs in all branches across the genome, repeated this 10,000 times and recorded the proportion of times substitutions were observed in  $\geq x$  resistant and  $\leq y$  sensitive branches. This proportion serves as an empirical  $P$  value for an unexpectedly high level of convergence among resistant branches, suggesting the action of selection. To be considered a candidate for positive selection, we required a SNP to have  $P < 0.05$  across all phylogenetic and ancestral reconstruction methods.

As multiple different SNPs within the same gene might nevertheless code for similar resistance phenotypes, we expanded the convergence test beyond individual SNPs to include whole genes and intergenic regions. In this method, branches were defined as convergent if they contained a SNP in the same gene or region, even if the SNPs occurred at different nucleotide positions. For each gene and intergenic region in the genome, we counted the number of SNPs occurring within the region boundaries, counting at most one SNP for each branch and counting SNPs in order of their frequency in the phylogeny. Using the same empirical resampling strategy as for SNP-based convergence, we generated a list of significant region-based convergence among resistant



branches. The genes found to be under selection using the phylogeny-based convergence method are detailed in **Supplementary Table 5**. Details on the pairwise convergence test and the analysis of the density of resistance-specific mutations are given in **Supplementary Figure 8**, **Supplementary Tables 14** and **15**, and the **Supplementary Note**.

**Selection testing by first-line drug phenotype.** PhyC and other supplementary tests for selection were performed similarly for resistance to each of the five first-line tuberculosis drugs: isoniazid, rifampicin, ethambutol, pyrazinamide and streptomycin (**Supplementary Tables 10** and **16–19**). The genes found to be significant by the broad resistance phenotype and resistance to isoniazid, rifampicin and streptomycin were highly similar with few exceptions. This similarity is likely a result of close associations among resistance phenotypes to isoniazid, rifampicin, ethambutol and streptomycin (for example, 82% of isolates resistant to either isoniazid or rifampin were resistant to both; **Supplementary Table 11**). The number of genes found to be significant for pyrazinamide was significantly lower than for the other drugs, likely because of the larger number of isolates with a missing resistance phenotype for this drug and the resultant low statistical power.

**SNPs detected in genes under selection.** All the SNPs seen in the targets of independent mutation in resistant isolates are listed in **Supplementary Table 20**. SNPs were called relative to the preceding (ancestral) node for each isolate in the phylogenetic tree. A multiple-sequence alignment of the genetic sequences for all SNP sites in the targets of independent mutation (including sites occurring in resistant strains only and in sensitive strains only and SNP sites occurring in both types of strains) is provided in **Supplementary Table 21**.

**Candidate gene variants in isolates with unexplained resistance.** We identified nonsynonymous SNPs in the genes under positive selection in isolates with unexplained resistance. We filtered out SNPs in these genes that occurred in any isolates sensitive to each drug. The results are detailed in **Supplementary Table 8**.

***M. tuberculosis* mutant generation.** Rv0050, encompassing *ponA1*, was replaced with a hygromycin resistance cassette using mycobacterial

recombineering in the H37Rv host strain. The *ponA1::hyg* replacement was confirmed by PCR and whole-genome sequencing. As we sought to identify mutants more likely to be independently causative of resistance, we focused on *ponA1* SNP alleles c.123C>G and c.1095G>T, as these occurred in mostly drug-resistant clinical strains, whereas alleles at a third site (c.1891C>T) were more prevalent in susceptible strains. SNP alleles in *ponA1* were generated by site-directed mutagenesis and were confirmed by Sanger sequencing. Wild-type or SNP alleles in *ponA1* were cloned under the control of a constitutive promoter and integrated in the *M. tuberculosis* genome as single copies at the L5 phage integration site.

**MIC assays.** All strains were grown in Middlebrook 7H9 medium supplemented with 0.25% glycerol, 10% oleic acid–albumin–dextrose–catalase and 0.05% Tween-80. For MIC calculations, the strains  $\Delta ponA1$ ,  $\Delta ponA1$  L5::*ponA1*<sub>wild-type</sub>,  $\Delta ponA1$  L5::*ponA1*<sub>C123G</sub> and  $\Delta ponA1$  L5::*ponA1*<sub>G1095T</sub> were grown until mid-log phase (0.5–0.8 spectroscopic optical density at 600 nm (OD<sub>600</sub>)) and then diluted to a calculated starting OD<sub>600</sub> of 0.006 and grown with or without drug for 6 d at 37 °C with shaking. Growth under all conditions was performed in duplicate. Two sets of duplicate experiments were performed at slightly different drug concentrations. Percent survival was calculated by normalizing the OD<sub>600</sub> measurement for each strain to that of its respective untreated control. MIC was defined as the drug concentration that inhibited growth to the extent that it was 1% or less of that observed for the untreated control.

40. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
41. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
42. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
43. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
44. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
45. Popescu, A.-A., Huber, K.T. & Paradis, E. ape 3.0: new tools for distance based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537 (2012).